DOCUMENT RESUME

ED 403 309 TM 026 041

AUTHOR Haertel, Edward H.
TITLE Latent Traits or Latent States? The Role of Discrete Models for Ability and Performance.
PUB DATE Apr 92
NOTE 32p.; The Raymond B. Cattell Award Invited Address presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992). For computer programs related to this study, see TM 026 042.
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; Data Analysis; *Error of Measurement; Generalizability Theory; *Item Response Theory; *Mathematical Models; Outcomes of Education; *Performance Factors; Scoring; Test Interpretation; Test Theory; Test Use
IDENTIFIERS *Latent Structure Models; *Mapping

ABSTRACT

Classical test theory, item response theory, and generalizability theory all treat the abilities to be measured as continuous variables, and the items of a test as independent probes of underlying continua. These models are well-suited to measuring the broad, diffuse traits of traditional differential psychology, but not for measuring the outcomes of school learning. Discrete latent structure models offer a powerful and promising alternative. Abilities can be modeled as partially ordered sets of discrete states (at a minimum, "nonmastery" and "mastery") and may be linked according to an asymmetric "prerequisite" relation. Narrower, simpler abilities may be combined into broader, more complex abilities. The various possible outcomes of performing a task can be modeled as a partially ordered set of task performance states. Abilities and task performances are clearly distinguished from one another, and more than one ability pattern may permit successful performance of a given task. Subtasks need not be modeled as conditionally independent given ability. The mapping from ability states to task performance states shows clearly what a given test can and cannot measure, and what may be inferred from a given pattern of test performance. These models for ability and task performance, together with the mapping between them, may be augmented with a suitable model for measurement error (misclassification) to complete an alternative framework for scoring, analyzing, and interpreting test performance. This framework has the potential to solve significant measurement problems inherent in performance testing and other applications. (Contains 12 figures.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
EDWARD H. HAERTEL
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 403 309

Latent Traits or Latent States?
The Role of Discrete Models for Ability and Performance

Edward H. Haertel
Stanford University

Raymond B. Cattell Award Invited Address, Presented at the Meeting of the American Educational Research Association, April 1992, San Francisco, California

TM026041

BEST COPY AVAILABLE

Latent Traits or Latent States?
The Role of Discrete Models for Ability and Performance

Edward H. Haertel
Stanford University

Abstract

Classical test theory, item response theory, and generalizability theory all treat the abilities to be measured as continuous variables, and the items of a test as independent probes of underlying continua. These models are well suited to measuring the broad, diffuse traits of traditional differential psychology, but not for measuring the outcomes of school learning. Unlike the items on an objective test, which may be regarded as statistically independent of one another, the subtasks of "real-world" tasks often depend on successful completion of prior subtasks. That is one reason why scoring and analyzing performance tests using traditional models is proving problematical.

Discrete latent structure models offer a powerful and promising alternative. Abilities can be modeled as partially ordered sets of discrete states (at a minimum, "nonmastery" and "mastery"), and may linked according to an asymmetric "prerequisite" relation. Narrower, simpler abilities may be combined into broader, more complex abilities.

The various possible outcomes of performing a task can be modeled as a partially ordered set of task performance states. Under this conception, abilities and task performances are clearly distinguished from one another, and more than one ability pattern may permit successful performance of a given task. Unlike the items of tests conforming to the usual IRT assumptions, subtasks need not be modeled as conditionally independent given ability.

The mapping from ability states to task performance states shows clearly what a given test can and cannot measure and what may be inferred from a given pattern of test performance.

These models for ability and task performance, together with the mapping between them, may be augmented with a suitable model for measurement error (misclassification) to complete an alternative framework for scoring, analyzing, and interpreting test performance. This framework has the potential to solve significant measurement problems inherent in performance testing and other applications.

Latent Traits or Latent States?
The Role of Discrete Models for Ability and Performance

Edward H. Haertel
Stanford University

These are interesting times for educational research, and especially for educational measurement. Jason Millman observed at an AERA meeting a couple years ago that the 70s had been the decade of criterion-referenced testing and the 80s of item response theory, and predicted that the 90s would be the decade of performance testing. So far, it looks as though he's right. Having failed to solve all the problems of our educational system by mandating more CRTs, minimum competency tests, teacher competency tests, and other multiple-choice instruments, legislators and educational policy makers concluded, not that "more testing won't solve our problems," but rather, that "we need a different kind of test." More telling, a majority of educational researchers, measurement specialists, and classroom teachers also seem to concur that we've been relying too heavily on objective tests. The consensus seems clear: We do need new kinds of tests. Evaluating schools, curricula, and learning outcomes on the basis of students' actual performance of meaningful tasks makes a lot of sense. (That's not to say, of course, that any kind of testing, in itself, is enough to reform U.S. education, but performance testing may well contribute to the solution of many educational problems.)

Performance testing poses major new challenges to the philosophy as well as the technology of educational measurement. Just building these new tests will be difficult, but even that alone will not be enough. We will need to think about and talk about intended learning outcomes in different ways; and we will need to develop new measurement models to score, analyze, and report test performance.

I'm going to begin this morning by describing a major difficulty with the straightforward application of classical test theory, item response theory, or even generalizability theory to performance test data. After that, I'll talk some about a different conception of ability and task performance, which leads naturally to methods of data analysis using latent class models, and describe some work David Wiley and I are doing, applying these models to item response data. I'm afraid all the examples I have to show you today use multiple-choice item responses, but even so, I think I'll be able to demonstrate the importance of thinking about student abilities in different terms if we are to do justice to the potential of performance testing for educational improvement and reform.

The major difficulty I see in applying classical test theory, IRT, or G-theory to performance test data is related to what may be called the granularity of performance tests. Responses from performance tests come in big chunks. Each performance exercise takes much longer to administer than an objective test item, and so an entire assessment necessarily consists of only a few exercises, often only a single one. This is a problem because, as stated by the well-known Spearman-Brown prophesy formula, reliability increases with test length. All of our major psychometric tools are designed for use with tests that consist of a fairly large number of small, separate pieces of evidence about the respondent. We generally assume that separate item responses are conditionally independent given examinee ability; in other words, that the only reason test items correlate with one another is because they depend on the common, underlying ability or abilities the test is designed to measure. This "conditional independence" assumption is important. It means that we cannot, for example, treat each step of a multi-step word problem as a separate test item, because students who fail earlier steps may never have a chance to demonstrate whether or not they could have accomplished later steps.

A performance exercise generally sets forth much more information for the student to take in than does a multiple-choice question; even more information than the reading passage that might precede a set of several multiple-choice questions. One reason for presenting more information is to provide more realistic contexts for students to demonstrate what they have learned. Another is to enable the measurement of higher-order thinking. In order for students to demonstrate their ability to, for example, integrate large amounts of information, sort out the relevant from the irrelevant, or apply what they know to solve real-world problems, the statement of the problem must be more complex than a multiple-choice question. Some performance exercises may also present more information because by design, they require students to interpret pictures, graphs, or other sources of information. Student responses are also more elaborate than their responses to objective tests. They may write extended responses, keep notebooks, or demonstrate their use of equipment. The fact that performance exercises present more information and elicit more elaborate responses has a couple of consequences for measurement. Most obviously, a performance exercise takes a lot longer to administer than a single test item. In addition, even if in the course of carrying out the performance exercise a student produces a series of separate, scorable responses, these are unlikely to be conditionally independent due to their common dependence on the same problem stimuli as well as the student's prior actions. (I note in passing that violations of conditional independence due to common stimuli could arise in multiple-choice reading comprehension tests where several

items depend on the same passage, but there, test developers usually work at writing items that are as independent of one another as possible.)

The most straightforward way, then, to apply our familiar psychometric models to performance test data would be to treat each entire exercise as a single item. That is, each exercise would be scored dichotomously, as "pass" or "fail"; a large number of performance exercises would be given to each student; and an examinee's total score for the assessment would be the number of exercises passed. If each exercise were treated as a separate unit, then the conditional independence assumption would be satisfied.

When an "exercise" is a laboratory activity, writing task, or open-ended problem that may require anywhere from ten or fifteen minutes to several days or more, this simply won't do. Some way has to be found to wring more than a single bit of information from such an extensive sample of respondent behavior. An easy step in that direction would be to rate overall performance on each exercise using, say, a six-point scale. We know that a well-designed item with ordered response categories can yield about as much information from each response threshold as a single binary item. Thus, an exercise scored on a six-point scale, with five thresholds separating the successive response categories, might yield as much information as about five dichotomous items. But if the exercise takes several orders of magnitude more time to administer, this is still not nearly enough information to justify the time required.

For their own instructional purposes (not just grading but also evaluating curriculum or instruction; identifying individual students' strengths, interests, or learning difficulties; instructional grouping or pacing), teachers will probably continue to get along without any more sophisticated psychometric models. It's not clear that these instructional purposes have <u>ever</u> been very well served by our measurement theory. Teachers will continue to observe and note, form tentative hypotheses and check them out, use anecdotal evidence to capture and communicate their understandings of the students in their care.

The granularity of performance exercise responses poses much more serious problems in the context of testing programs that must summarize the performance of many schools or students in a common metric, on a common scale. If no individual scores are provided, as with the National Assessment of Educational Progress, for example, then matrix sampling provides a feasible, if costly, solution--each student can be given only one or a few exercises, and the responses of different students to different exercises can be summarized to characterize the overall performance of a population of

examinees across a domain of exercises. This solution keeps down the testing time for each individual respondent, but it does not address the fundamental inefficiency of taking a large block of student time and generating only a single score. Moreover, matrix sampling is of little help if comparable scores are required for individual examinees. In the short run, most testing programs with mandates both to include performance testing and to provide individual scores are introducing a few performance exercises, but continuing to rely mostly on objective test questions to attain acceptable levels of reliability, at least until more and better performance exercises can be developed. That may work for now, but it seems likely that within a few years, better methods for scoring the performance exercises themselves will have to be found.

What might these better methods be? I believe that to find an answer, we must begin with a reconsideration of the structure of the abilities we are trying to measure, and of the measurement tasks we use to gather information about them.

The notion of "ability" in classical test theory, IRT, etc. is pretty fuzzy. The ability is "whatever the test measures." It is usually modeled as a unidimensional continuum, a single scale along which different individuals (and often test items) can be arrayed. Multidimensional extensions of this basic model are essentially variants of a Thurstone factor model: Two or more different abilities are posited, which can be mixed like the colors of an artist's palate to form the particular composite ability representing degrees of proficiency on some given task. The relation of abilities to one another is nearly always expressed as a correlation, telling no more than the degree to which higher levels of one ability tend to be associated with higher levels of another. It is implicit in this model that having more of one ability can compensate for having less of another. Note also that the correlation coefficient is <u>symmetric</u>. Neither ability X nor ability Y comes first, they just go together. These models were developed to measure broad human abilities, the traits of differential psychology, like G_f or G_c. The boundaries of these abilities are indistinct, and the span of tasks to which they apply is enormous.

This conception of abilities as broad, unidimensional continua symmetrically associated with one another is quite <u>unlike</u> the conception of ability implicit in our organization of the school curriculum. There, we treat different abilities as <u>ordered</u>--more basic capabilities are taught before more advanced ones, new learning building on what has been learned before--and as <u>specific</u>--the particular skills taught and learned in school each are applicable to a fairly

well defined set of tasks. Moreover, abilities are not generally substitutable. If a child is having difficulty with a math word problem that calls for reading, setting up an equation, and carrying out some calculation, the skillful teacher will not assume that improving any one of those three abilities sufficiently will remove the difficulty.

(I should say in passing that I recognize the risk in overstating the sequential nature of the curriculum. Even at the early grade levels, work in reading, mathematics, and so forth should be meaningful and should involve "higher-order" thinking. I don't mean to imply that rote "tool skills" must be well learned before the beginning of any "meaningful" application. Moreover, different children can and will solve the same problem in different ways, capitalizing on their particular strengths and compensating for their particular weaknesses. But that being said, if multiplication is taught in a way that entails the addition of partial products and addition is taught in a way that does not entail any multiplication, then some facility with addition is prerequisite to learning multiplication, and not the other way around.)

What I've described so far seems to be something of an anomaly. How can it be that we've done as well as we have taking measurement models devised for broad human traits, things like "verbal ability" or "quantitative reasoning", and using them to measure school achievement? And what does any of this have to do with performance testing? I believe that we've done as well as we have with classical test theory and IRT because we've artificially limited ourselves to an extremely simplified and unrealistic task structure, one in which there are no logical dependencies, no prerequisite relations, among what David Wiley and I would refer to as the subtasks! We limit ourselves to that kind of task when we insist that items, the subtasks of a test, be conditionally independent given ability, and that no item require the prior solution of an earlier item.

This is a major reason why objective tests have been found so unsatisfactory by educators and policy makers alike--in the real world, subtasks are connected! And one of the major reasons performance tests pose such an important challenge to our psychometric models is that, unless they are simply treated as indivisible units to be passed or failed or rated on some ordered scale, their complex internal structure must be acknowledged.

I believe that performance tasks can be constructed with an ample number of distinct, scorable units. But those units, i.e. subtasks, will depend on one another in complex ways. Two important differences between conventional test items and performance subtasks are first, that carrying out a given

subtask may require other subtasks to be completed or attempted first, and second, that most complex problems can be completely and correctly solved without even attempting, let alone correctly performing, all of the different subtasks that might be part of one or another path to a solution.

I began this morning by saying that I'd first sketch some problems I see with the application of most of our current measurement models to performance test data, and then set forth an alternative conception and give some examples. I'd like to turn now to that second topic, and take a few minutes to sketch an alternative view of abilities and task performances that I believe is more than adequate to characterize the kinds of subtask dependencies and alternative solution paths that will occur in complex performance exercises. The formal model is set forth in a chapter David Wiley and I have written for a forthcoming book, Test Theory for a New Generation of Tests, edited by Norm Fredericksen, Bob Mislevy, and Isaac Bejar.

Insert Figure 1 about here

I've just been talking about tasks and subtasks, but now I'm going to turn to the structure of the underlying abilities that make it possible to perform different tasks. That's an important distinction. Briefly, an ability is modeled not as a continuum, but as a collection of two or more distinct states. The simplest model is one with exactly two states, not possessing the ability or possessing it, or, if you like, nonmastery and mastery. More complex structures include intermediate states (denoting different patterns of partial mastery), which are partially ordered. Note that in the middle figure, for example, the intermediate states (0,1) and (1,0) are not ordered--neither of these is lower or higher than the other. Learning, in these models, is represented by transitions from one ability state to another, always between two states connected by an arrow. Reaching the "1" state indicates full mastery. In the second illustration, full mastery could be reached by either of two different paths, each including just one of the two intermediate states. The final illustration shows an ability with three distinct states. This structure is called a chain, and has the properties that there is only one path from the initial, pre-instructional state to the final state, and that there is an order relation between any two states of this ability. Much more complicated structures are possible.

Insert Figure 2 about here

Another very important feature of this model is that abilities may be combined to form more complex structures, which also meet the definition of single abilities. This next transparency shows two dichotomous abilities, **a** and **b**,

and the three different ways they could be combined into a single, more complex ability. The top box shows the structure of the two separate abilities. Each has two states, **0** and **1**. The next box shows the way the two combine if **a** is prerequisite to **b**, in other words, if mastery of ability **a** is a necessary precondition for mastery of **b**. In this case, the two separate dichotomous abilities could be treated as a single ability with three states.

In our theoretical development, David Wiley and I model each ability as a mathematical structure called a distributive lattice. This a collection of discrete performance states that are partially ordered and satisfy a few other specific conditions. A collection of separate abilities (each one itself a collection of performance states) forms a partially ordered set, or poset. It can be shown that, given sensible rules for combining abilities, the set of distinct states an individual might occupy with respect to all of the abilities taken together also forms a distributive lattice. That's what I illustrated in this last transparency. It is what enables us to combine simple abilities into more complex abilities. This means, for example, that in principle, we could incorporate the relatively narrow instructional goals of a series of separate learning units and the broader goals of the year's course work within a single framework, showing explicitly how learning outcomes at different levels of generality and accomplished in different amounts of time were related.

As I said before, it is important to distinguish between task performances and the underlying abilities that enable them. I think this is one of the places where criterion-referenced testing missed the mark in the 1970s--Formulating "behavioral objectives" led us to identify underlying abilities too closely with particular manifest performances. Most educationally significant abilities, especially the so-called "higher-order" abilities, should be relevant to a range of disparate tasks. As we move into performance testing, we may again be tempted to define intended learning outcomes in terms of the performance of specific tasks. Maintaining a clear conceptual distinction between abilities and task performances will be critical if we want performance testing and the instructional practices it encourages to take us beyond low-level recapitulation of learned procedures. Just getting students to do particular tasks will not be enough.

I'm going to turn now from abilities back to tasks again. In our work, Wiley and I have defined tasks as goal-directed activities, bounded in time, for which one or more outcomes can be evaluated. Tasks are generally composed of subtasks, which may be related in different ways. Superficially, the structures of tasks look a little like the ability structures I've shown you, with nodes representing subtasks and

different kinds of arrows showing how they're related, but there are important differences between the mathematical structures of tasks versus abilities.

A given subtask may be solved in different ways, but for each subtask, there is at least one configuration of abilities (or more accurately, configuration of ability states) enabling its successful performance, and at least one configuration that does not enable its successful performance. Thus, a subtask induces a partition of the set of all possible ability states into those enabling and those not enabling its performance. Given an interrelated set of subtasks, each with its ability requirements, it is possible to determine exactly what inferences about ability can be made from different patterns of subtask performance.

I've talked now about abilities and about tasks, but I must briefly describe one more piece of the puzzle, before turning to some examples. That final piece is a method of accounting for measurement error. In this framework, measurement error includes two different sources of discrepancies between the record of an individual's task performance and the task performance state that would be predicted from that individual's underlying abilities. First of all, people don't always perform in a manner that accurately reflects their underlying capabilities. Lapses of attention, failures of motivation, careless errors may all lead to task performances that fall short of what would be predicted from the actual profile of underlying abilities. Likewise, lucky guesses, inadvertent hints, or faulty solution procedures that happen to work for particular problems may lead to manifest performances that exceed what would be predicted from the true abilities. Second, in addition to the problem of people not performing in a way that reflects their true capabilities, the record of the performance may not accurately reflect what actually occurred. For my purposes today, there's no need to disentangle these sources of error any further. In multiple-choice items, I model measurement error using either one or two parameters for each item. An item's "false positive" parameter gives the probability that an examinee who does not, in fact, possess the requisite abilities will nonetheless produce a correct response to the item. A "false negative" parameter gives the probability that an examinee who does possess the requisite abilities will respond incorrectly. In some cases, I'll assume that the false negative probability is zero, and model only false positives.

One big difference between classical test theory or the continuous latent structure models of IRT versus the discrete latent structure models I've been describing is in the amount of attention that must be paid to the specific structure of the problem. These are generally not "generic" models that

can be applied without any tailoring or modification to any old test.

It's remarkable, sometimes, how little IRT and factor models require one to know about what one is testing. Several years ago, I had a paper accepted in which I had reanalyzed one of those "classic" data sets that appear in the literature and are reanalyzed by many different authors. These were the data from the Law School Admissions Test, or "LSAT", first published by Bock and Lieberman in 1970. Ivo Molenaar, the editor of the journal, wisely asked me whether I might obtain the actual LSAT items from which the original data were obtained and see whether my interpretation of the data, that there were two distinct types of items, was supported. In the course of tracking down the actual test, I found that even though several leading scholars had used these data in first-rate papers introducing major new analytical techniques, none of these authors had ever had occasion to actually inspect the items themselves. Bock and Lieberman's published table of the number of examinees giving each possible response pattern was all they needed, and all I'd used in the first draft of my paper. I found myself stumbling in trying to explain to a very helpful and well intentioned secretary at the Law School Admissions Council how it could be that so many psychometricians had studied the test and published so many papers about it without ever having seen it!

Insert Figure 3 about here

Let me show you a couple different analysis of the LSAT data I just told you about. This figure shows the varimax-rotated factor loadings for items 11-15 from Section 7 of the LSAT. This is essentially the solution published by several authors who have used these data to illustrate computational methods for the factor analysis of dichotomized variables. It is well accepted that two factors are required to account for the pattern of associations among the five items. Two things are evident from this figure. First, items 12 and 13 load most heavily on factor 2, while items 11, 14, and 15 load most heavily on factor 1. Second, the communalities of items 11 and 13 are markedly higher than those of the remaining three items. (Because the axes are orthogonal, the communality is simply the sum of the squares of the two factor loadings for a given item. On the figure, it's also the square of the distance from the origin to the point plotted for an item.) This two-factor solution suggests that the five items each require a different mixture of the two underlying abilities. Look at item 11, for example. Because .798 is about five or six times .139, it appears that in solving item 11, the ability represented by the first factor is five or six times as important as the ability represented

by the second. For item 14, the ratio is only about 1 1/4 to one.

Insert Figure 4 about here

Now let me show you the results of fitting a latent class model to the same data. In this model, there are two dichotomous abilities, which I've labeled 1 and 2. You'll notice that these two abilities define three ability patterns: Neither 1 nor 2; 1 only; or 1 and 2. Items 11, 14, and 15 require only ability number 1, whereas Items 12 and 13 require ability 2. The model doesn't distinguish whether Items 12 or 13 also require ability 1 or not. Because everyone who has the second ability must also have the first, exactly the same examinees will be able to solve these items whether or not they require the first ability. The top part of the table shows the false positive and false probabilities estimated for each item, and the lower part shows the proportions of examinees with each combination of abilities. All in all, this model requires estimation of about as many parameters as the continuous model, and provides about as good a fit. In the figure at the top of the transparency, I display this solution in a form analogous to the continuous solution.

When I actually examined the items, I <u>did</u> find major differences between items 12 and 13 versus 11, 14, and 15 in the structure of the items and the processes required for their solution. I did <u>not</u> find any reason why items 11 and 13 should have had such high communalities relative to the other items, which gives me an additional reason for preferring my discrete solution to the continuous solution.

Let me present just one more illustrative analysis, this one not yet published, using the ten core test items in physics from the second IEA science study. These are all multiple-choice items, each with five response alternatives. The data are from the United States population 2(B), N = 2,519, which I treated as a simple random sample. I'd like to describe the steps I went through in analyzing the data in some detail, but I should say in advance that I'm still experimenting with these methods, and the details are still in flux.

My method of looking at the data focuses on particular patterns of responses across the ten items. I assume that there are some small number of "real" patterns, what I'll refer to as <u>latent response patterns</u>, each of which corresponds to some set of ability patterns that might be found in the group of examinees tested. If there were no measurement error, no false positives or false negatives, then every examinee's actual responses, their <u>manifest response patterns</u>, would be identical to one of these latent response patterns. In practice, of course, because of

measurement error, examinees may guess the right answer to questions they can't really solve, so not all of the manifest response patterns found in the data correspond to possible latent response patterns.

In order to discover the latent response patterns, I wanted to remove the effects of measurement error from the data. For this analysis, I assumed that there were no false negative errors, just false positives, so I needed to estimate just one misclassification parameter for each of the ten items, namely its false positive probability. If I knew in advance what the latent response patterns were, it would be straightforward to get maximum likelihood estimates simultaneously of both the proportions of respondents conforming to each latent response pattern and the false positive probabilities for each item, using a program like Cliff Clogg's MLLSA. (I've run models for ten items and over 50 latent classes on a Macintosh Powerbook 170.) In this case, though, I didn't know in advance what the latent response patterns were. In order to estimate the misclassification probabilities without knowing the latent response patterns, I took all possible subsets of 4 items from among the ten physics items, and fit the same model to each of those 210 four-item subsets. (There are 210 ways of choosing 4 items from a set of 10.) I was pretty sure that the model I chose included all of the latent response patterns to just those four items. If there were some extra latent response patterns, for my purposes that didn't matter. To fit these 210 models, I arranged the items in each four-item set in order of decreasing p-value, and fit a model with the twelve latent response patterns shown on the next transparency.

Insert Figure 5 about here

With four items, there only are 16 possible latent response patterns, and my models included twelve of them. The only patterns <u>not</u> included were: being able to solve only the most difficult of the four items and none of the three easiest (that would be the pattern "0001"); only being able to do the second most difficult (that would be "0010"; only the two most difficult ("0011"); or only the three most difficult ("0111"). With 12 latent classes and 4 misclassification probabilities, you might expect these models each to fit perfectly--counting parameters suggests that they should be just identified. In fact, the fits were very good, but not perfect--The models are what Goodman referred to as "pseudo-identified." The parameter values required to reproduce the data exactly sometimes fell slightly outside of the range from zero to one, and so the corresponding parameter estimate would go to the boundary. In particular, I expected, and found, that many of the latent class proportions were estimated to be zero. That was fine. Putting an extra

latent response pattern into the model and having the proportion in that class go to zero was essentially the same as not putting it in at all, and would not affect the estimates of the false positive probabilities, which were all I cared about at this stage. Each item was included in 84 of these 210 analyses, so I got 84 separate estimates of its false positive probability. (After choosing one item, there are 84 ways of choosing 3 more from among the remaining 9.) I sorted these 84 separate estimates, constructed stem-and-leaf diagrams, and used the stem-and-leafs to determine the final estimated false positive rate for the item.

Insert Figure 6 about here

Item 6 illustrates how this worked. Originally, I'd intended to simply take the median of the 84 estimates for each item, but things turned out not to be quite that simple. From the latent response patterns in the previous transparency, you'll recall that patterns are included for examinees able to solve only the easiest item (pattern 1000) or only the second easiest item (pattern 0100). Item 6 happens to be the fifth easiest item, with four easier and five harder. That meant that in the 84 runs involving item six along with some three others, item six was the easiest or second easiest 50 times, and was the hardest or second hardest 34 times. Thus, there were 50 runs that included a latent response pattern for examinees able to solve item 6 and none of the other items. These 50 runs all yielded estimated false positive rates between .151 and .169, with a median of .158, so .158 was used as the false positive rate for item 6. The remaining 34 models did not include such a class, and turned out to be misspecified. Because they did not include one of the actual latent response patterns, they yielded biased estimates of the false positive probability for item 6. This kind of split, with markedly higher false positive estimates resulting from runs that did not include a latent response pattern for ability to solve the target item only, was found only for the easier items--there was no evidence for the two or three most difficult items that any examinees could solve one of them and no other items.

Insert Figure 7 about here

The false positive probabilities for the ten items ranged from .126 to .242, with a median of .164. There's little evidence of any correlation with item difficulty. Note that under this model, observed correct responses may represent either actual knowledge of the answer to the item, or a false positive by an examinee who in fact does not know the answer. Given the observed p-value and the conditional probability of such an errorful response, it's easy to calculate the underlying proportion of examinees who actually possess the abilities required to solve the item. These are the "true" (or latent) p-values in the rightmost column. You'll note

that because of the exceptionally high false positive probability for item 3, it has the lowest true p-value of any of the ten items, even though it's only ninth in order of observed p-value.

Assuming that misclassifications occur independently for each item, it is straightforward to calculate the probability of an examinee conforming to any possible latent response pattern producing any possible manifest response pattern. All these probabilities can be arranged in a 1024 by 1024 matrix. By inverting that matrix and premultiplying it by the vector of observed response pattern frequencies for the 1024 response patterns, it is straightforward to reproduce the vector of latent response patterns. (Fortunately, the 1024 by 1024 matrix is the Kronecker product of ten two-by-two matrices, one for each item, so its inverse can be calculated just by inverting the ten two-by-two matrices and taking the Kronecker product of the inverses.) The next transparency shows the effect of this operation on the distribution of response pattern frequencies.

Insert Figure 8 about here

You can see that the effect is to "concentrate" the examinees in fewer high-frequency response patterns than before. The top 30 observed response patterns accounted for only 37 percent of all the examinees, but the top 30 latent response patterns accounted for 62 percent. I should point out that these are different patterns. The most frequent observed pattern (90 students) was 1100011100, i.e., getting items 1, 2, 5, 6, and 7 correct. The pattern represented by ten zeros, getting all the items wrong, was eleventh on the list. In the latent response patterns, the pattern of all zeros was at the top of the list, and the 1, 2, 5, 6, 7 pattern was second. There are other rearrangements throughout. In the observed data, 532 of the 1024 possible patterns occurred, although over 200 of those were produced by only one examinee. Estimated frequencies were nonzero for all 1024 of the latent response patterns, but the great majority had very small or slightly negative frequency estimates.

Insert Figure 9 about here

Here are all the latent response patterns with estimated frequencies of 25.000 or higher. I've rearranged the items in order of decreasing p-value, from easiest to hardest. I also inserted a space between the first five items and the last five, to highlight an apparent break between these two sets of items. You'll notice that with the exception of the two patterns I've marked with an asterisk, all these patterns show either partial mastery of the first five items and complete nonmastery of the second five, or else complete mastery of the first five items and partial mastery of the second five. To put it differently, except for the starred

patterns, no one in these high-frequency patterns is able to do any of the last five items unless they're able to do all of the first five. Based on an inspection of the actual items, I believe that the first five items each can be solved on the basis of general reasoning and out-of-school experience, whereas the last five also require specific knowledge that students would be unlikely to obtain outside of formal coursework in science.

Insert Figure 10 about here

Let me focus in on the structure of these two separate sets of items. You'll recall that all ten items define 1,024 possible latent response patterns, but focusing on just five items at a time, there are only 32 latent response patterns. This next transparency shows the estimated frequencies of all those patterns for the first five and for the last five items. There's a striking difference apparent here. For the first five items, nearly all of the possible latent response patterns occur with significant frequencies, whereas for the last five items, there are a few high-frequency patterns and a good number of patterns that are very rare, even some with estimated frequencies less than zero! This difference is reflected in the relative complexity of the lattice structures required to represent these high-frequency patterns.

Insert Figure 11 about here

For the easy items, the ones that I hypothesize may be answered on the basis of out-of-school learning, I required three separate structures to accommodate all of the highest-frequency patterns. (These structures must follow certain rules, which I describe in the chapter David Wiley and I wrote for the forthcoming book by Fredericksen, et al.) What this says is that individual items may be solved in more than one way, or by using more than one mix of abilities. The sample of examinees represents a mixture of students at different points within these different structures. The difficult items, which I hypothesized to be more closely tied to formal schooling, show a cleaner, simpler structure. This makes sense. Informal, out-of-school learning is less systematic than school course work, and it is reasonable to suppose that students might pick up the particular abilities or bits of information required to answer different items more-or-less independently of one another. Consequently, there are a lot of different knowledge patterns manifested. The harder items, which I have suggested are more closely tied to the school curriculum, show a more linear, sequential structure, and just one lattice suffices to show all of the high-frequency latent response patterns. This may mean that these items are each more likely to be solved in just one way, and it also means that students are more likely to

acquire the abilities to solve these items in a more-or-less fixed order.

My final transparency shows an alternative analysis of these items, based on all of the latent response patterns rather than just the high-frequency patterns. To construct this table, I first constructed two-by-two tables for all possible pairs of items, showing the estimated numbers of examinees actually able to solve neither item, only the first, only the second, or both. I expressed the numbers in all these four-fold tables as proportions, then took the numbers in the two off-diagonal cells of each four-fold table and arranged them in a single, large table.

Insert Figure 12 about here

Once again, I've arranged the items in order of decreasing p-value, from easiest to hardest. The proportion of examinees estimated to be able to answer item x correctly and not item y appears in row x, column y. For example, just over twelve percent of the examinees, .124, could solve item 2 and not item 8. About 17 percent could solve item 8 and not 2. I've put in dashed lines separating the easiest five items from the hardest to make the table easier to read, but also to highlight their structure. Like the lattices I just showed you, this table shows sharply defined prerequisite relations between many pairs of items, but a linear ordering, a Guttman scale, is not complex enough to capture it. The upper left quadrant of the figure shows that taking any two of the first five items, in either order, a substantial fraction of the examinees, from about 10 to 25 percent, can solve the first and not the second. The lower right quadrant shows that the second five items are more strongly ordered. Within this quadrant, the values below the diagonal are substantially smaller than those above. The upper right and lower left quadrants of the figure confirm the strong ordering between the domains represented by the easier versus the more difficult items. There are a lot of values in the lower left quadrant that are very near zero, suggesting that the abilities required by the column item are a subset of those required by the row item. I should point out that a table constructed in this way using the observed response patterns rather than the estimated latent response patterns would show similar trends, but would be much less sharply defined. The underlying structure is revealed much more sharply after the correction for guessing, which enabled me to focus on underlying, latent response patterns rather than observed (manifest) response patterns.

I hope that these examples have given you some idea of the potential of discrete latent structure models to reveal more about the meaning of task performance than do the continuous models now in general use. Let me say again that I would

have liked to present an analysis of some performance test data, so that I could show you how these methods can accommodate more complex task structures, with logically determined, prerequisite relations between pairs of subtasks instead of the artificial independence of a collection of test items. I'm presently pursuing work with performance test data, funded by the National Science Foundation, and I hope to have more to report at AERA next year in Atlanta.

Performance tests will be built and used whether or not more adequate methods are developed for analysis and reporting. But unless such methods are developed, I'm afraid that the performance testing movement may fall far short of its potential to inform and improve education. It's happened before. The criterion-referenced testing movement of the 1970s offered the educational research community a chance to explore much more deeply the meaning of criterion-referenced test interpretations, the substantive meaning of standards, and methods of standard setting anchored empirically to students' real-world performances. Important scholarly work was done on these issues, of course, but by and large, educational practice was limited to the "80 percent correct" standard, with interpretations driven more by the names given to tests than by any serious study of the underlying abilities their items elicited. The challenge I see for the 1990s is to develop and demonstrate methods of analyzing, summarizing, and reporting performance test data that fulfill the promise of really showing what students know as well as what they can do.

References

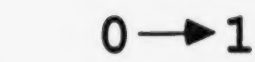
Haertel, E. H. (1990a). Continuous and discrete latent structure models for item response data. Psychometrika, 55, 477-494.

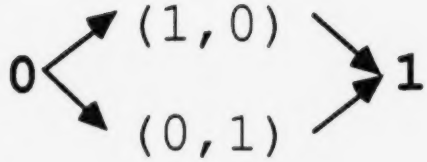
Haertel, E. H. (1990b). Using restricted latent class models to map the skill structure of achievement items. Journal of Educational Measurement, 26, 301-321.

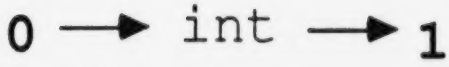
Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislevy, and I. Bejar (Eds.), Test theory for a new generation of tests (pp. 359-384). Hillsdale, NJ: Erlbaum.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), Improving inquiry in social science (pp. 75-107). Hillsdale, NJ: Erlbaum. (Also as Studies of Educative Processes, No. 20, Northwestern University, 1987.)

Wiley, D. E., & Haertel, E. H. (1996). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In M. B. Kane & R. Mitchell (Eds.), Implementing Performance Assessment: Promises, Problems, and Challenges (pp. 61-89). Hillsdale, NJ: Erlbaum.

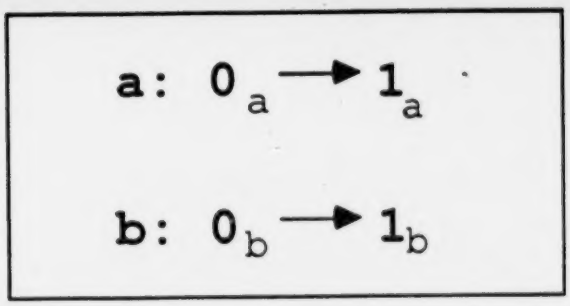


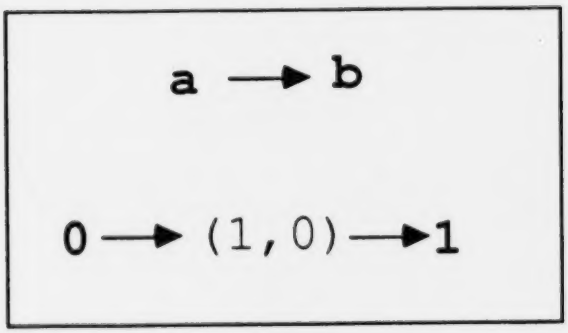


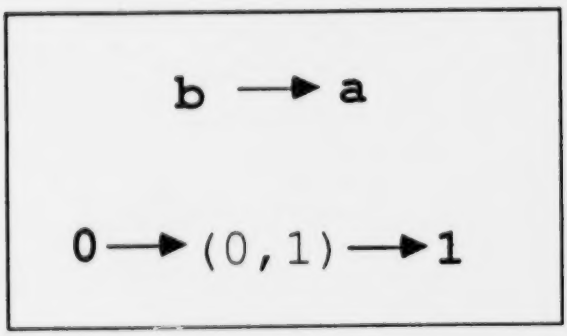


Examples of Discrete Abilities

Figure 1


a: $0_a \rightarrow 1_a$
b: $0_b \rightarrow 1_b$


a → b
0 → (1,0) → 1


b → a
0 → (0,1) → 1

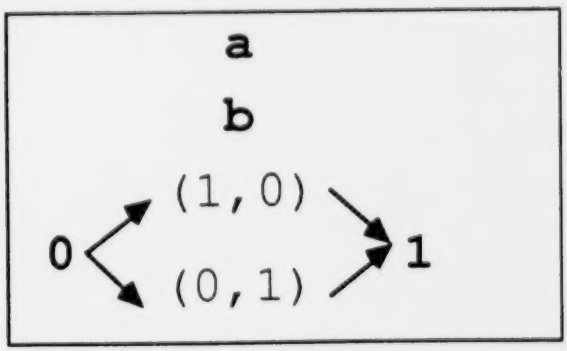
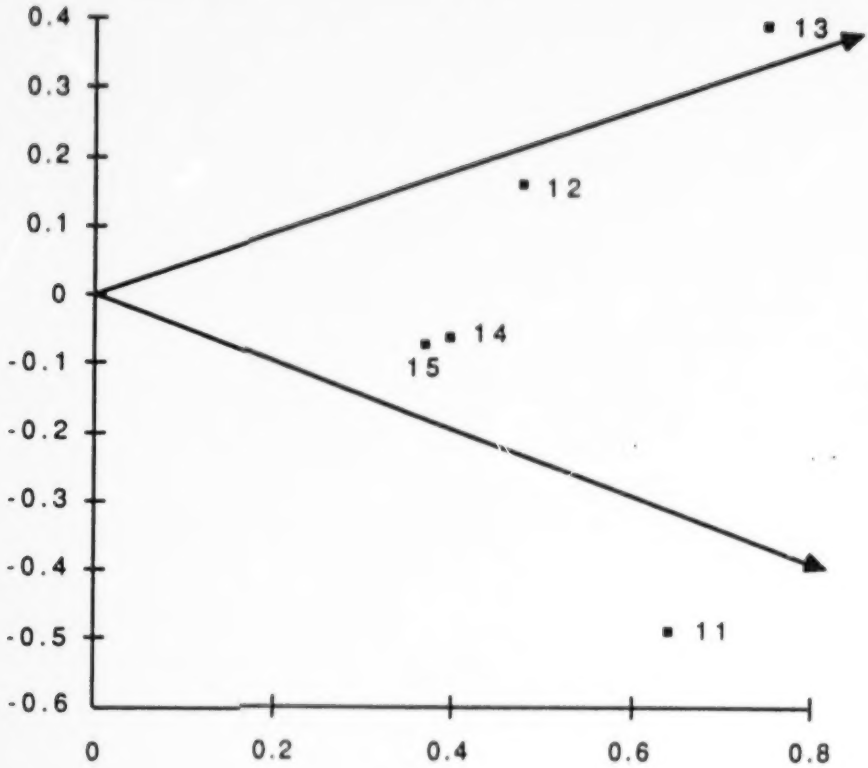

a
b
(1,0)
0
(0,1)
1

Figure 2



Two-Factor Solution for LSAT-7 Items

Figure 3

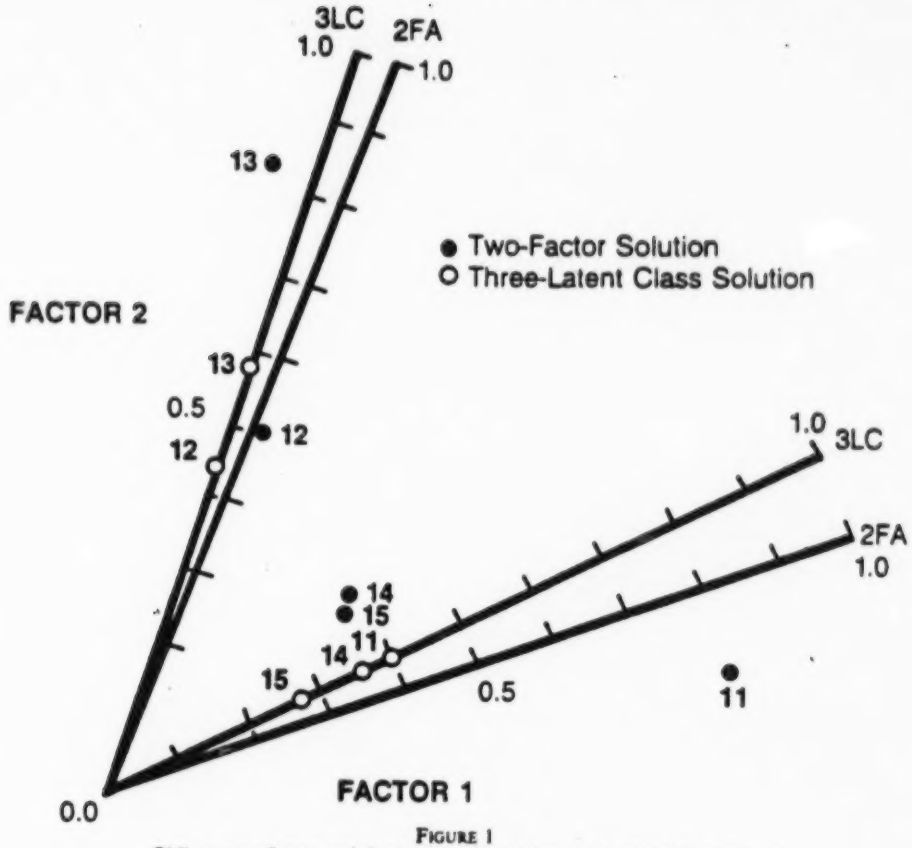


FIGURE 1
Oblique two-factor and three-latent class solutions for LSAT7 items.

Item	Abilities Required	False Positive Probability	False Negative Probability
11	1 only	0.5383	0.0915
12	1 and 2	0.3793	0.1859
13	1 and 2	0.4482	0.0466
14	1 only	0.2679	0.3000
15	1 only	0.6506	0.1035

Proportions of Examinees With Different Abilities	
Neither 1 nor 2	0.2175
1 only	0.1415
1 and 2	0.641

Figure 4

Four-Item Models Used to Estimate
False Positive Probabilities

Latent Response Patterns*
0000
1000
0100
1100
1010
1001
0110
0101
1110
1101
1011
1111

*Items in order of decreasing p-value, i.e., least to most difficult.

Figure 5

Stem-and-Leaf of False Positive Rate Estimates for Item 6

```
.13|
.14|
.15|1134455555556666666677788889
.16|00000111112233334688 99
.17|
.18|
.19|
.20|
.21|
.22|
.23|
.24|
.25|
.26|
.27|
.28|
.29|
.30|
.31|4455
.32|111335
.33|
.34|
.35|
.36|899
.37|003
.38|79
.39|00
.40|033468
.41|5679
.42|8
.43|11
.44|6
.45|
.46|
```

Figure 6

Statistics for IEA Physics Items

item	false positive rate	Observed p value	"True" (latent) p value
8	.163	.783	.741
2	.166	.743	.693
1	.166	.727	.673
7	.163	.696	.636
6	.158	.650	.586
10	.139	.417	.323
5	.126	.332	.236
9	.185	.319	.165
4	.139	.197	.067
3	.242	.275	.044

Figure 7

Highest 30 Observed and Latent Response Pattern Frequencies

	Manifest Response Patterns				Estimated (latent) response patterns		
	frequency	proportion	cumulative percent		frequency	proportion	cumulative percent
1	90	0.0357	3.57		194.393	0.0772	7.72
2	74	0.0294	6.51		168.509	0.0669	14.41
3	50	0.0198	8.50		102.900	0.0408	18.49
4	43	0.0171	10.20		79.470	0.0315	21.65
5	41	0.0163	11.83		76.488	0.0304	24.68
6	41	0.0163	13.46		73.967	0.0294	27.62
7	37	0.0147	14.93		65.624	0.0261	30.22
8	35	0.0139	16.32		60.325	0.0239	32.62
9	34	0.0135	17.67		55.005	0.0218	34.80
10	34	0.0135	19.02		54.419	0.0216	36.96
11	32	0.0127	20.29		48.058	0.0191	38.87
12	32	0.0127	21.56		44.690	0.0177	40.65
13	31	0.0123	22.79		42.692	0.0169	42.34
14	29	0.0115	23.94		42.395	0.0168	44.02
15	26	0.0103	24.97		40.686	0.0162	45.64
16	26	0.0103	26.00		40.079	0.0159	47.23
17	26	0.0103	27.03		35.320	0.0140	48.63
18	25	0.0099	28.03		34.089	0.0135	49.98
19	23	0.0091	28.94		31.412	0.0125	51.23
20	23	0.0091	29.85		30.729	0.0122	52.45
21	21	0.0083	30.69		29.336	0.0116	53.62
22	21	0.0083	31.52		26.346	0.0105	54.66
23	20	0.0079	32.31		26.093	0.0104	55.70
24	19	0.0075	33.07		24.436	0.0097	56.67
25	18	0.0071	33.78		24.333	0.0097	57.63
26	18	0.0071	34.50		24.013	0.0095	58.59
27	18	0.0071	35.21		23.647	0.0094	59.53
28	17	0.0067	35.89		23.116	0.0092	60.44
29	17	0.0067	36.56		22.839	0.0091	61.35
30	16	0.0064	37.20		22.773	0.0090	62.25

Figure 8

Estimated Latent Response Pattern Frequencies ≥ 25.0 for 10 IEA Items, Ordered by Difficulty

1	194.393	00000 00000	
2	168.509	11111 00000	
3	102.900	11111 10000	
4	79.470	11110 00000	
5	76.488	11101 00000	
6	73.967	11111 01000	
7	65.624	11000 00000	
8	60.325	10111 00000	
9	55.005	11011 00000	
10	54.419	01100 00000	
11	48.058	10010 00000	
12	44.690	10111 10000	*
13	42.692	11100 00000	
14	42.395	11111 01100	
15	40.686	11111 11000	
16	40.079	11110 10000	*
17	35.320	11111 00100	
18	34.089	01111 00000	
19	31.412	10001 00000	
20	30.729	11010 00000	
21	29.336	10110 00000	
22	26.346	11111 10100	
23	26.093	11111 11110	

* Response pattern for which at least one of first 5 items is incorrect and at least one of last 5 items is correct.

Figure 9

Latent Response Pattern Frequencies for
IEA Physics Items 8, 2, 1, 7, 6

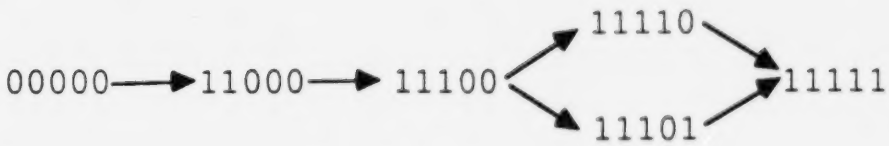
1	663.639	11111
2	224.729	11110
3	187.227	00000
4	168.223	11101
5	118.793	10111
6	101.437	11011
7	99.664	11100
8	76.245	01111
9	62.462	11000
10	60.925	11010
11	58.604	10010
12	56.923	10011
13	56.366	10110
14	52.799	10101
15	50.528	11001
16	48.009	00100
17	45.652	01110
18	41.924	01000
19	40.796	01100
20	36.570	10001
21	32.612	10100
22	32.093	01011
23	28.722	01101
24	26.467	00010
25	25.985	00111
26	24.442	01010
27	22.939	01001
28	22.396	10000
29	20.045	00001
30	18.759	00011
31	11.572	00110
32	1.452	00101

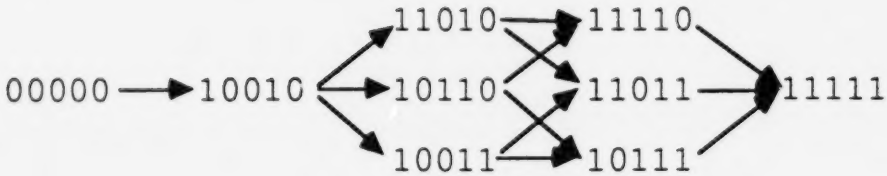
Latent Response Pattern Frequencies for
IEA Physics Items 10, 5, 9, 4, 3

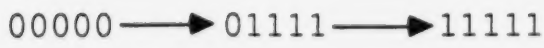
1	1166.733	00000
2	442.282	10000
3	230.748	01000
4	132.413	11000
5	105.869	00100
6	78.720	01100
7	54.034	10100
8	41.619	11110
9	39.729	10001
10	35.470	11100
11	25.164	00110
12	25.072	10010
13	24.411	01010
14	24.090	10110
15	22.544	00001
16	18.232	10101
17	15.738	01110
18	14.555	11011
19	13.669	00101
20	11.672	01001
21	5.298	11010
22	4.964	01101
23	4.498	11111
24	3.182	01011
25	2.956	00011
26	1.307	00010
27	0.773	11101
28	-1.295	01111
29	-1.392	00111
30	-5.446	10111
31	-8.255	11001
32	-10.350	10011

Figure 10

Latent Response Pattern Lattices for First Five IEA Items, Ordered by Difficulty (8, 2, 1, 7, 6)







Latent Response Pattern Lattice for Last Five IEA Items, Ordered by Difficulty (10,5,9,4,3)

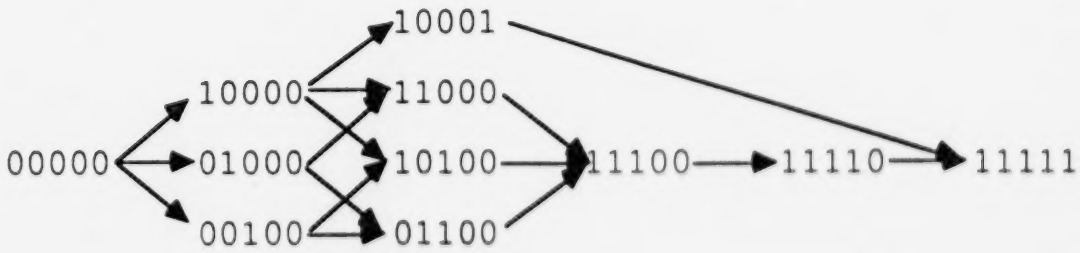


Figure 11

Estimated Proportions Able to Answer <Row> and Not <Column> Correctly

Item	8	2	1	7	6	10	5	9	4	3
8		0.173	0.179	0.209	0.245	0.469	0.545	0.592	0.677	0.708
2	0.124		0.158	0.205	0.238	0.450	0.513	0.552	0.634	0.651
1	0.111	0.138		0.187	0.222	0.421	0.493	0.508	0.613	0.638
7	0.104	0.148	0.151		0.202	0.396	0.468	0.489	0.573	0.604
6	0.090	0.132	0.135	0.151		0.361	0.420	0.459	0.521	0.560
10	0.051	0.080	0.071	0.083	0.098		0.233	0.254	0.284	0.302
5	0.040	0.056	0.056	0.067	0.070	0.146		0.164	0.193	0.224
9	0.016	0.024	-0.001	0.018	0.038	0.096	0.093		0.124	0.151
4	0.003	0.008	0.007	0.004	0.002	0.028	0.024	0.026		0.065
3	0.011	0.002	0.009	0.012	0.018	0.022	0.032	0.030	0.065	

Highest-Frequency Latent Response Patterns

freq.	8	2	1	7	6	10	5	9	4	3
194.39	0	0	0	0	0	0	0	0	0	0
65.62	1	1	0	0	0	0	0	0	0	0
54.42	0	1	1	0	0	0	0	0	0	0
79.47	1	1	1	1	0	0	0	0	0	0
76.49	1	1	1	0	1	0	0	0	0	0
55.01	1	1	0	1	1	0	0	0	0	0
60.33	1	0	1	1	1	0	0	0	0	0
168.51	1	1	1	1	1	0	0	0	0	0
102.90	1	1	1	1	1	1	0	0	0	0
73.97	1	1	1	1	1	0	1	0	0	0

Figure 12